Hybrid Machine Learning-Based Air Quality Forecasting Using CatBoost with Hunger Games Search and Arithmetic Optimization Algorithm

Xiaowen Geng
School of Engineering, Guangzhou College of Technology and Business, Guangzhou City, Guangdong Province 510000 China
E-mail: mannmann0504@126.com

Keywords: environment, air pollution, ML, air quality index, catboost

Recieved: January 18, 2025

Air pollution is a significant global concern, posing a major challenge to sustainable development if neglected. Leveraging mathematical frameworks through ML offers an optimal and cost-effective solution for modeling air pollution. This investigation introduces hybrid ML-based frameworks to anticipate air quality pollutants and classify air quality. Specifically, the CatBoost algorithm was combined with the Arithmetic Optimization Algorithm (AOA) and the Hunger Games Search algorithm (HGS) for prediction and classification purposes. The database comprises daily time series data of air pollutants in China from 2018 to 2021. Autocorrelation function (ACF) and partial autocorrelation function (PACF) approaches were utilized to select input combinations for each pollutant. Results indicate that the integrated model provides highly accurate forecasts of pollution index time series using the regression method. Furthermore, evaluation metrics reveal that the classification method surpasses the regression method regarding accuracy for predicting the AQI.

Povzetek: Članek predstavlja hibridna modela CatBoost-AOA in CatBoost-HGS za napovedovanje onesnaženosti zraka na osnovi časovnih vrst, kjer se uporabljata tudi optimizacijska algoritma HGS in AOA.

1 Introduction

Air pollution implies the existence of additives on natural compounds to the degree that they are caused by natural phenomena or air-polluting activities by humans. It implies the existence of any substance in the air that can be harmful to humans or their environment. Pollutants, whose number reaches more than 180, may be natural or man-made and exist in different forms such as solid particles, liquid droplets, or gas [1,2].

Its basic resources are industries and factories, as well as vehicles. The gradual and long-term but silent effects of air pollution have made officials and people pay less attention to it. This is while the upward trend of death statistics, cancer, and heart attacks caused by air pollution displays that many industrialized societies have suffered a gradual death, the main cause of which is air pollution. Air pollution is, therefore, fatal to humans in the long run [3,4].

Air pollution is a worldwide problem that severely impacts human health and the environment as well. As many as 4 million premature deaths occur every year because of air pollution, according to reports from the WHO [5]. Due to this issue, air quality monitoring has turned out to be highly significant these days. Real-time monitoring allows early warnings of air-polluting events that are so crucial to protecting public health [6]. This article describes the development of a web-based air quality warning system that uses streaming data in support of dynamic environmental management. Traditionally, air quality has been monitored by periodic measurements of the levels of air pollution. However, with advanced technology, air pollution levels can now be continuously monitored in real time. Real-time monitoring is thus capable of providing early warnings of air pollution events, which are one of the critical premises for public health protection [6]. Real-time monitoring can also enable the capture of extensive amounts of data, whereby trends and patterns of air pollution levels may be disaggregated over time [7].

The increasing demand for accurate air quality forecasting models has led to the development of hybrid machine learning frameworks that integrate multiple algorithms to enhance predictive performance. This study aims to address the following key research questions:

How can hybrid machine learning models improve the accuracy of air quality predictions compared to traditional statistical methods?

Which optimization techniques (such as AOA and HGS) yield the best performance when combined with CatBoost for air quality forecasting?

What is the impact of feature selection and time-lagged variables on the predictive capabilities of air pollution models?

How does the proposed hybrid framework compare to benchmark models in terms of computational efficiency and real-time applicability?

An appropriate air quality predictive model is to be developed, which will help in developing an effective early warning system. The model can be developed through various techniques, including the statistical model, ML model, and neural network model. These

statistical frameworks are simple to develop and not responsive in capturing the complexity of the variables of air pollutants and other atmospheric factors [8]. The ML frameworks of relatively higher complexity and able to capture non-linear interaction between the variables are RFs and SVMs. Last but not least, NN frameworks, to be more precise, coming under the category of DL, are an even more complex model; hence, they can detect very subtle patterns in data [9].

It will allow the implementation of early warning systems and mitigation measures that avoid harmful impacts on public health and the environment. During the last years, several research works were undertaken for the development of various approaches of air pollutant forecasting. The next section reviews the mainstream approaches to air pollutant forecasting.

- ML-based approaches:

These ML-based approaches have been of more recent interest because they can model complex nonlinear relationships between meteorological and air pollutant variables. Several ML techniques, such as ANN, SVR, and RF, have been utilized for the prediction of air pollutant density. One of the very latest publications is that of Song et al. (2023), where they used the XGBoost algorithm, based on ML, for forecasting the density of PM2.5. It was found that the use of XGBoost was effective in optimizing the hyperparameters of the model [10].

- Time series-based approaches:

The time series data-based approaches essentially make their forecasting of future air pollutant densities based on historical data. The approaches based on time series are widely applied owing to their simplicity and their capability for the portrayal of seasonal and temporal variations in air pollutant densities. Bhatti et al. (2021) have used the SARIMA model to anticipate the PM2.5 density. It followed from the given analysis that the SARIMA model had high prediction accuracy and was suitable for time series anticipation of PM2.5 density [11].

- Hybrid approaches:

Hybrid approaches combine at least two frameworks to boost the precision of forecasts of air pollutants. Very recently, Liu and Chen recommended a neural network-based hybrid framework for PM2.5 density anticipation whose performance outperforms that of an individual model [12].

Also, Kaushal et al. (2025) proposed a hybrid CatBoost-SVR model for earthquake prediction using the LANL dataset, achieving superior accuracy over individual models. Their findings highlight the effectiveness of hybrid ML approaches in complex forecasting tasks, aligning with our study's methodology for air quality prediction [13].

- Statistical frameworks:

Various statistical frameworks have been utilized to anticipate the density of air pollutants, which include linear regression, multivariate regression, and principal component regression. A very recent study applies a multivariate LR framework for PM2.5 density anticipation in urban areas. As explained in this investigation, the model showed very high accuracy in its prediction [14].

- Physical frameworks:

The physical frameworks, in turn, simulate the physical processes governing dispersion using mathematical equations. Recently, Kong et al. (2021) have used the WRF-Chem model for forecasting PM2.5 densities within China. The WRF-Chem model presented accuracy within the forecast [15].

In other words, the selection of an optimal approach for forecasting air pollutants may be heavily dependent on several parameters; besides, it is often common to combine more than one approach to come out with more accurate and reliable predictions. However, hybrid frameworks using DL and ML frameworks, PCA-based linear regression frameworks, and time series-based frameworks like SARIMA are some prominent approaches that have displayed promising results in recent research.

In the last years, genuinely promising results on AQI prediction using ML techniques, based on several environmental factors, have been possible to observe. Some of the works analyze the use of classification and regression approaches for AQI forecasting. An example is the work of Li et al., recently published in the Journal of Cleaner Production, where an SVM classifier and an RF regression framework are recommended to anticipate AQI in China. Other studies have estimated the AQI using a combination of ANN and multiple linear regressions by Shams et al. (2023) [16].

Other research has also been conducted on the effectiveness of the utilization of machine learning (ML) approaches in AQI prediction. For example, Jiang et al., in their study published in Atmospheric Environment, compared the productivity of many ML frameworks, including LR, DT, and ANNs, to anticipate the AQI in Beijing, China. In another study published in the Journal of Environmental Management, Sharma et al. applied multiple linear regression and ANNs to meteorological variables and road traffic data to anticipate AQI in Jaipur, India.

ML classification approaches have also received broad attention in air quality forecasting due to their competence in the forecast of air pollutant densities and improvement of air quality management. These provide a proper and efficient way of analyzing not only the complicated relationship between air quality parameters and meteorological variables but also identifying key factors affecting air quality. Over the last few years, several articles have appeared on the state-of-the-art applications of ML classification approaches for air quality forecasts in prestigious journals.

The investigation adapted the ML classification approaches to simulate AQI using meteorological variables and air pollution densities based on an Atmospheric Environment publication. In the investigation, SVR, RF, and neural network frameworks were applied for the simulation of AQI in Beijing, China, with great precision. A similar study in the Journal of Environmental Management applied ML classification approaches for predicting air pollutant densities concerning traffic and meteorological variables in the urban area. Decision tree, RF, and k-nearest neighbor frameworks were employed in this investigation to anticipate the density of particulate matter, nitrogen

oxides, and sulfur dioxide. These studies therefore hint at the potential of ML classification approaches in air quality forecasting and their capability to provide efficient and accurate predictions. Given the ever-increasing demand for effective air quality management, applications of such approaches are foreseen to increase further in the coming years.

Some studies showed that the approaches of regression and classification performed well, but their success depended on the database and modeling techniques adopted. This paper compares and evaluates regression and classification frameworks based on their performance in AQI prediction. In the present study, the considered model is the CatBoost model, linked to 2 metaheuristic algorithms, applied to a database of AQI measurements combined with environmental factors. The remainder of this paper is outlined below:

In Section 2, materials and approaches are described. In the continuation of this section, while explaining the methodology, the algorithms used in the research, evaluation indices, and research database are explained. Section 3 reviews the results of the case study and discusses them. Finally, the conclusion is displayed in Section 4.

Table 1 presents a comparative analysis of various air quality forecasting approaches based on recent research. It categorizes different methodologies, including machine learning, time series models, hybrid frameworks, statistical techniques, and physical simulation models. This comparison provides insights into the strengths of different models and their effectiveness in predicting air pollutant levels and AQI.

Table 1: Comparison of air quality forecasting approaches in recent studies

Key Findings	Evaluation Metrics	Dataset	Authors (Year)	Algorithms Used	Approach
XGBoost was effective in optimizing hyperparameters and predicting PM2.5 levels.	Hyperparameter tuning	PM2.5 data	Song et al. (2023)	XGBoost	**ML-Based Approaches**
SARIMA demonstrated high accuracy and suitability for time series-based PM2.5 forecasting.	Forecast accuracy	Historical PM2.5 data	Bhatti et al. (2021)	SARIMA	**Time Series-Based Approaches**
The WRF-Chem model provided accurate PM2.5 forecasts for China.	Model accuracy	Air pollution data in China	Kong et al. (2021)	WRF-Chem	**Physical Frameworks**
SVM classification and RF regression were effective for AQI forecasting in China.	Prediction accuracy	AQI data from China	Li et al. (2020)	SVM (classification), RF (regression)	**AQI Prediction with ML**
Combining ANN and MLR produced reliable AQI forecasts.	Prediction accuracy	AQI data	Shams et al. (2023)	ANN + Multiple Linear Regression	**Hybrid AQI Prediction Models**
ANN-based models performed better than LR and DT in AQI forecasting.	Model comparison	AQI data from Beijing	Jiang et al. (2019)	LR, DT, ANN	**ML Model Comparisons**

2 Materials and approaches

2.1 Methodology

The purpose of this investigation is to provide an approach to boost the precision of daily anticipation of air pollutants based on the capabilities of hybridization and ML approaches. The series of data related to air quality has random, irregular, and unstable characteristics, and this has made it difficult to anticipate pollutants and air quality.

In the following, the methodology of the research will be explained. The first step is data collection. The database used in this investigation includes seven pollutants related to air quality, including SO2, NO2, O3, PM10, PM2.5, and AQI, all in µg/m3 units, as well as CO pollutants in mg/m3 units. After data collection, pre-processing takes

place. Preprocessing is an efficient step in boosting the accuracy of predictions and reducing the processing time of special algorithms in AI. In this investigation, the prediction of each air pollutant is based on the historical data of that pollutant. In other words, predictions are made based on a function of values at earlier time steps (Q1, Q2, ...). To determine the various input combinations, the ACF and PACF of the observed data are used. There is a temporal dependence between the observations in the time series discussion. The phrase "autocorrelation" refers to "serial correlation," or the dependency between sequence values across time, since statistics frequently refer to correlation as dependence. The ACF is the function that determines autocorrelation as a function of the time interval between observations. The ACF was applied to identify significant lag dependencies, helping to determine the overall structure of autocorrelations. PACF was then utilized to select the appropriate lag order for modeling the autoregressive (AR) component of the time series. This approach ensured that only the most relevant lagged variables were included in the forecasting models. ACF and PACF are used to determine the process's time- or location-specific behavior in random processes and time series.

Following this, with the help of statistical appraisal indexes and by using the method of the SVR, it selected the best inputs. Supervised learning algorithms are for SVM intended to be used for classification and regression analysis. The key importance of SVMs is that they can handle high-dimensional data with excellent performance; it is excellent for working in cases where sample sizes are small, which means that they can be versatile for many applications. In application fields boasting its capacious ability to precisely identify complex features and categories, the SVM algorithm is employed. After choosing the best input combinations for each pollutant separately, two hybrid frameworks, CatBoost-HGS and CatBoost-AOA, were formed for prediction and classification with the CatBoost algorithm and AOA and HGS algorithms. The ACF/PACF-based input selection process was specifically applied to the regression models, ensuring that only the most significant lag variables were used. In contrast, the classification task utilized AQI class labels derived from the original AQI values. Furthermore, the two approaches—regression and classification—were compared independently using appropriate evaluation metrics: RMSE and MAE for regression, and Accuracy, Precision, and F1-Score for classification. This comparison highlights the suitability of classification models, such as CatBoost-HGS, for AQI forecasting, as classification effectively captures the discrete categories of AQI levels. the incorporation tries to boost and adjust the CatBoost hyperparameters at every stage of processing. Accuracy in the pollutant prediction of recommended hybrid frameworks was considered by different evaluation indices, taking into account error analyses in what follows. Other analyses, such as sensitivity analysis and correlation between variables, also formed a part of this investigation.

2.2 Evaluation indices related to regression

The evaluation indices employed in this investigation to compare the precision of regression frameworks include MBE, RMSE, MAPE, R^2, JSD, and RAE, expressed drawing on the subsequent formulas:

$$MBE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (2)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (4)$$

$$JSD(P \,||\, Q) = (KL(P \,||\, M) + KL(Q \,||\, M)) / 2 \quad (5)$$

$$RAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|y_i - \bar{y}|} \quad (6)$$

where N displays the count of observations, y_i and $\hat{y}_i$ are ith real value and ith estimated value, respectively, and $\bar{y}$ displays the mean of data points, $JSD(P \,||\, Q)$ displays Jensen-Shannon divergence between P and Q, and $M = (P + Q) / 2$.

The evaluation indices used in this investigation related to classification include Precision, Recall, and F1 Score. Based on potential states for existing and projected samples, 4 probable states—TP (true positive), TN (true negative), FP (false positive), and FN (false negative)—will probably occur. These 4 factors are used to derive the evaluation indices using the following equations [17,18]:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (8)$$

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (9)$$

2.3 Categorical boosting (CatBoost)

CatBoost is a gradient-boosting ML framework that is particularly effective for handling categorical features. It is an open-source library developed by Yandex and offers high performance and accuracy in various ML tasks. CatBoost provides seamless handling of categorical variables so that explicit preprocessing involving one-hot encoding is not explicitly required. There is an internal implementation of a novel algorithm, "Ordered Boosting," which uses statistical approaches and gradient boosting intelligently to handle categorical features effectively. CatBoost comes under the gradient boosting framework, where an ensemble of weak decision trees is built sequentially. It iteratively improves the model by fitting new trees to the residuals of the previous iterations, intending to minimize a loss function. It has built-in mechanisms for handling missing values in data. It learns from missing values either as another category or using the most frequent or mean value strategy. Model interpretability is supported by CatBoost because it allows both calculating and visualizing feature importance. These

are very helpful for understanding various features' influences on the model's predictions. In contrast to traditional boosting methods, CatBoost utilizes Ordered Boosting, which improves the handling of categorical features and reduces overfitting. This technique involves generating permutations of the data to maintain the temporal ordering of data points while ensuring unbiased updates to the model. The key advantage of Ordered Boosting is its ability to better capture the relationships in categorical features by considering their sequential dependencies, leading to more accurate predictions. This method also helps mitigate the common issue of overfitting seen in other gradient boosting algorithms, particularly when working with small datasets or highly imbalanced classes. Due to its robustness, high performance, and efficient handling of categorical variables, it gained its momentum in various ML competitions and real-world applications. It facilitates classification and regression; thus, this algorithm finds its applications in a wide range of areas, starting from customer churn prediction, recommendation systems, fraud detection, and many more [19,20].

2.4 SVR

SVR is an ML framework used for regression tasks. SVR is an extension of SVM for regression problems. In SVR, the goal is to find a function that approximates the mapping between input variables (features) and the continuous output variable. The algorithm aims to diminish the error between the projected and actual values while also adhering to a specified margin of tolerance. SVR requires labeled training data, consisting of input features and corresponding continuous target values. It is important to scale the input features to ensure they have similar ranges. SVR uses kernel functions to transform the input features into a higher-dimensional space. Commonly used kernels include linear, polynomial, radial basis function (RBF), and sigmoid. The SVR algorithm finds an optimal hyperplane in the transformed trait domain that boosts the margin while minimizing the error. The training process involves solving a constrained optimization problem. Once trained, the SVR model can be evaluated using suitable metrics. The trained SVR model can make predictions on new, unseen data by transforming the input features using the learned mapping and applying the regression function. SVR is a robust framework for regression tasks, especially when dealing with complex, non-linear relationships between features and targets. It is widely used in various domains, including finance, economics, and engineering, where predicting continuous values is essential [21–23].

2.5 HGS

HGS is a crowd-based framework that combines elements of competition and cooperation between individuals to search for optimal solutions. The main inspiration for this algorithm was the collaborative behavior of animals in nature to search for food when they are hungry. HGS is based on the actions and behavioral preferences of animals driven by hunger. This dynamic, fitness-wise search strategy makes the optimization process more comprehensible and consistent for novice users and decision-makers by basing it on the straightforward idea that "Hunger" is the most important homeostatic motivation and cause for all animal behaviors, decisions, and actions. To replicate the impact of hunger on each search stage, the HGS integrates the idea of hunger into the trait process. To put it another way, an adaptive weight drawing on the notion of hunger is produced and utilized. It adheres to the computationally logical rules (games) that nearly all animals play. These competitive games and activities are frequently flexible evolutionary strategies that elevate the chances of surviving and obtaining food. This approach is more efficient than the existing enhancement approaches because of its dynamic character, straightforward structure, high productivity regarding convergence, and acceptable quality of resolutions [24–26]. The reference to animals searching for food during hunger is intended to explain the biological inspiration behind the HGS algorithm. In nature, animals search for food based on their needs, often following certain patterns and behaviors. Similarly, the HGS algorithm imitates this natural process by searching for optimal solutions in a computationally logical manner. While the algorithm is inspired by nature, its design is mathematically structured to perform optimally in computational environments, balancing exploration and exploitation of the search space. The combination of biological inspiration and computational efficiency allows the algorithm to effectively solve optimization problems.

2.6 AOA

AOA is a population-based framework introduced in 2020 by Abualigah. The algorithm aims to optimize mathematical expressions by reducing redundant computations, simplifying expressions, or reordering operations to minimize the count of operations required. Some common arithmetic optimization approaches that are used include Constant Folding, Strength Reduction, Common Subexpression Elimination, Loop-Invariant Code Motion, Compiler optimization, etc. Constant folding involves evaluating constant expressions at compile time rather than runtime. This optimization replaces constant expressions with their computed values, eliminating unnecessary computations during program execution. Strength reduction replaces expensive operations with equivalent but cheaper operations. This technique reduces the computational complexity of arithmetic operations. Common subexpression elimination identifies and eliminates redundant computations by reusing the result of a previously computed subexpression. Instead of recomputing the same subexpression multiple times, the result is stored and reused when needed. Loop-invariant code motion moves computations outside loops when the result of the computation remains invariant throughout the loop iterations. By performing the computation once and reusing the result, redundant computations within loops can be avoided. Modern compilers employ various optimization techniques to optimize arithmetic operations

automatically. These include instruction scheduling, register allocation, loop optimizations, and more. Compiler optimizations take advantage of the specific hardware architecture and aim to generate efficient machine code for arithmetic computations. It's noteworthy that the choice and efficacy of the optimization approaches depend on factors such as the programming language, compiler, target platform, and specific characteristics of the arithmetic operations being optimized [27–29].

In the context of air quality forecasting, redundant computations in AOA are primarily associated with unnecessary re-evaluations of the objective function and excessive position updates within the population. This can lead to increased processing time without significant gains in solution accuracy. To address this, an adaptive convergence strategy has been incorporated to reduce redundant evaluations once the algorithm stabilizes, preventing excessive iterations. Additionally, dynamic step-size adjustments were introduced to refine the search process, minimizing ineffective position updates while maintaining the integrity of the optimization process. These refinements improve the computational efficiency of AOA, making it more suitable for large-scale environmental modeling where high-dimensional data processing is required.

In this study, hyperparameter tuning was performed using the Hunger Games Search (HGS) and Arithmetic Optimization Algorithm (AOA), which combine the exhaustive search capability of grid search with the computational efficiency of Hyperband's early-stopping strategy. This method allowed for the efficient exploration of a broad hyperparameter space while focusing resources on the most promising configurations. The tuning process was applied independently to both regression and classification tasks, with optimization based on 5-fold cross-validation performance on the training data.The optimal hyperparameters for the CatBoost model were found to be a learning rate of 0.1, a depth of 6, 500 iterations, an l2_leaf_reg of 3, random_strength set to 1, and early stopping rounds of 50.

algorithm 1 outlines the process of hyperparameter optimization using a hybrid approach that integrates CatBoost, Arithmetic Optimization Algorithm (AOA), and Hunger Games Search (HGS). The optimization process begins with initializing a population of random CatBoost hyperparameters. The AOA phase focuses on exploring and exploiting the search space through arithmetic operations, while the HGS phase refines the search by incorporating energy levels and "hunger" values for each solution. The algorithm iteratively updates the best solution based on performance evaluations and continues until the optimal hyperparameter set is found. This method efficiently balances exploration and exploitation, ensuring optimal tuning of hyperparameters for the CatBoost model.

Algorithm 1 Hybrid CatBoost-AOA-HGS Hyperparameter Optimization

Input: Dataset D, CatBoost model CB, AOA population size N, HGS iterations T
Output: Best hyperparameter set $H^\star$
Initialize AOA population $P = \{H_i\}_{i=1}^{N}$ with random CatBoost hyperparameters
Evaluate fitness $F(H_i) = \text{CatBoostEval}(CB, D, H_i)$ for each $H_i \in P$
Set best solution $H^\star = \arg\min_{H_i} F(H_i)$
for $t = 1$ to T **do**
    // AOA Phase
    for each $H_i \in P$ **do**
        Update exploration and exploitation parameters μ, MOA based on AOA rules
        Generate new candidate H_i' using arithmetic operations:
            $H_i' = H_i + \mu \cdot rand \cdot (H^\star - H_i)$ ▷ Example arithmetic update
        Evaluate $F(H_i')$
        if $F(H_i') < F(H_i)$ **then**
            $H_i \leftarrow H_i'$
        end if
    end for
    Update $H^\star$ if a better H_i is found
    // HGS Phase
    for each $H_i \in P$ **do**
        Compute energy level E_i and hunger level $Hunger_i$
        Select leader(s) based on minimum fitness
        Generate new solution H_i'' based on HGS operators:
            $H_i'' = H_i + rand \cdot (Leader - H_i) + Hunger_i \cdot rand$
        Evaluate $F(H_i'')$
        if $F(H_i'') < F(H_i)$ **then**
            $H_i \leftarrow H_i''$
        end if
    end for
    Update $H^\star$ if a better H_i is found
end for
return $H^\star$

Algorithm 1: Hybrid CatBoost-AOA-HGS Hyperparameter Optimization

2.7 Description of the database

The database used in this investigation includes a set of air pollutants collected in 2 major cities of China, i.e., Changchun, and Zhengzhou from 2018 to 2021. The dependent variable, or output, is the same as AQI. The input variables SO_2, NO_2, O_3, PM_{10}, and $PM_{2.5}$ are all in $\mu g/m^3$ units, as well as CO pollutants in mg/m^3 units. It should be noted that all data are collected on a daily basis. All air pollutants can be classified based on their chemical origin and physical state. These classifications are used to organize discussion and research in air pollution factors. Pollutants are divided into primary and secondary groups depending on their origin. Primary pollutants, such as carbon monoxide, sulfur dioxide, nitrogen oxides, and hydrocarbons, are those pollutants that have entered the atmosphere directly and are found in the atmosphere in the same way they are released. Secondary pollutants, such as ozone, are those pollutants that are formed in the atmosphere by a photochemical reaction as a result of hydrolysis or oxidation. Table 2 displays the statistical description of the research database. More information about the research database is available from reference [30]. In this table, individual pollutants (SO_2, NO_2, CO, O_3, PM_{10}, $PM_{2.5}$) represent direct atmospheric components, while AQI and Index serve as air quality indicators. AQI is a composite metric derived from pollutant concentrations, and Index categorizes AQI values into six discrete levels, representing different air quality conditions. Their inclusion in Table 2 provides a comprehensive overview of the dataset by connecting raw pollutant data with overall air quality assessments.

Table 2: Statistical description of pollutants and air quality indicators in the dataset

	Count	Mean	STD	Min	25%	50%	75%	Max
So2	8216	32.46762	35.22384	2	7	15	49	220
No2	8216	42.11648	22.68928	6	25	38	54	199
Co	8216	0.877057	0.507142	0.279	0.549	0.736	1.02425	6.5
O3	8216	61.14606	41.25921	3	29	53	84	232
Pm10	8216	103.3738	100.5177	6	48	78	124	1665
Pm2.5	8216	63.5611	74.02761	4	23	43	76	899
AQI	8216	96.27081	74.66038	16	53	73	111	500
Index	8216	2.324124	1.196457	1	2	2	3	6

In this database, the AQI index has 6 different labels from number 1 to 6, each of which displays air quality, which is displayed by the name of Index in Table 2. Unlike individual pollutants, AQI provides an aggregated measure of air pollution levels, and Index further classifies AQI into standard air quality categories. Their inclusion alongside pollutant statistics ensures a holistic representation of the dataset, facilitating a clearer interpretation of how pollutant levels correlate with air quality status. Therefore, drawing on the AQI label, air quality can be separated into 6 different groups. Table 3 displays the range values of AQI for each category, where each AQI range is associated with a specific air quality category to indicate pollution severity. The Category column provides a numerical encoding of these classifications, facilitating structured data processing in the study. This representation ensures consistency in air quality assessment and aligns with widely recognized air quality reporting frameworks. The classification structure allows for a clear interpretation of pollutant concentrations and their corresponding health implications. According to this table, with the increase of AQI values and placing the data in classes with a larger label, the amount of air quality decreases.

Table 3: Statistical description of the research database

AQI range	Description of the category	Category
0 to 50	Good	1
51 to 100	Moderate	2
101 to 150	Unhealthy for Sensitive Groups	3
151 to 200	Unhealthy	4
201 to 300	Very Unhealthy	5
301 to 500	Hazardous	6

To get familiar with the research variables, Fig. 1 displays the correlation matrix between the features. The dependent variable of AQI has a direct link with all independent (input) variables except O3. In the range of -1 to 1, correlation matrices display the link between 2 sets of data or variables. A correlation of -1 denotes a totally negative linear link, a correlation of 0 denotes no relationship at all, and a correlation of 1 denotes a positive linear relationship between 2 variables.

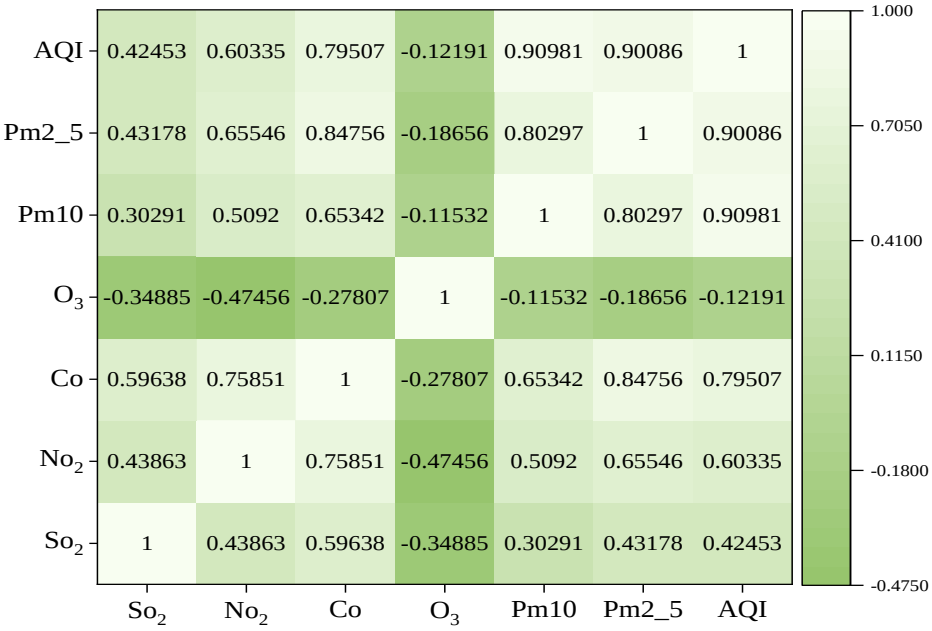


Figure 1: The correlation matrix between the features

As it is clear from this figure, Pm10 and Pm2.5 indices have the most positive correlation with AQI. In other words, with their increase, it can be expected to increase the value of AQI. Also, the correlation value between AQI and O3 indices is equal to -0.12191, which displays an inverse relationship between them.

Fig. 2 displays the sensitivity analysis of variables drawing on DMIM. The values of the delta and sensitivity indices in this graphic range from 0 to 1. The more closely these indices' values resemble 1, the more influence the pertinent variable has over the model's output. The delta index displays how much the distribution has changed, and the sensitivity index displays how much the output's variance has decreased [31]. This chart makes it evident that the PM10 variable has the most influence on the AQI. The sensitivity analysis indices for the PM2.5 variable have the greatest values after that. Conversely, because of its low indices, the O3 variable has the least effect on the AQI.

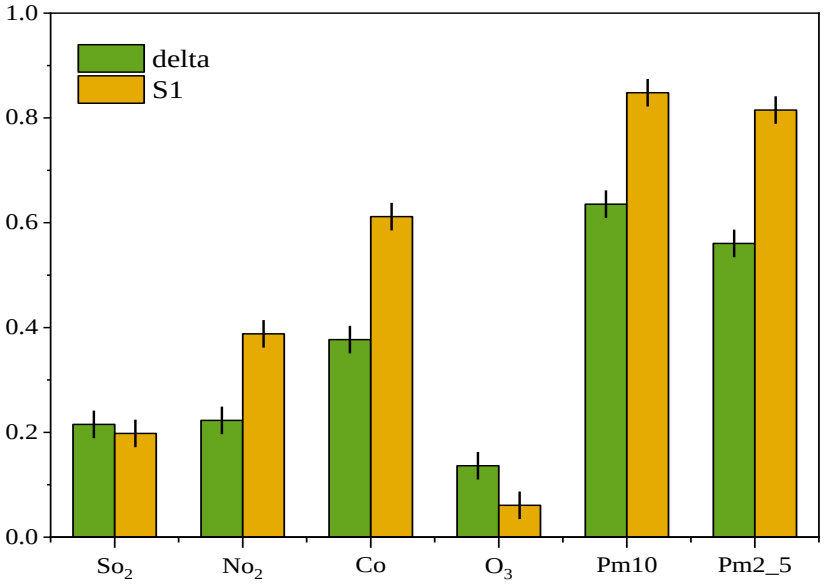


Figure 2: Sensitivity analysis of variables based on the DMIM method

3 Determining the Best Input Combinations

To find the best input combinations, the ACF and PACF of the observed data are used. In Fig. 3, the Lag-1–lag-25 ACF and PACF of the pollutants with the 95% confidence interval are displayed.

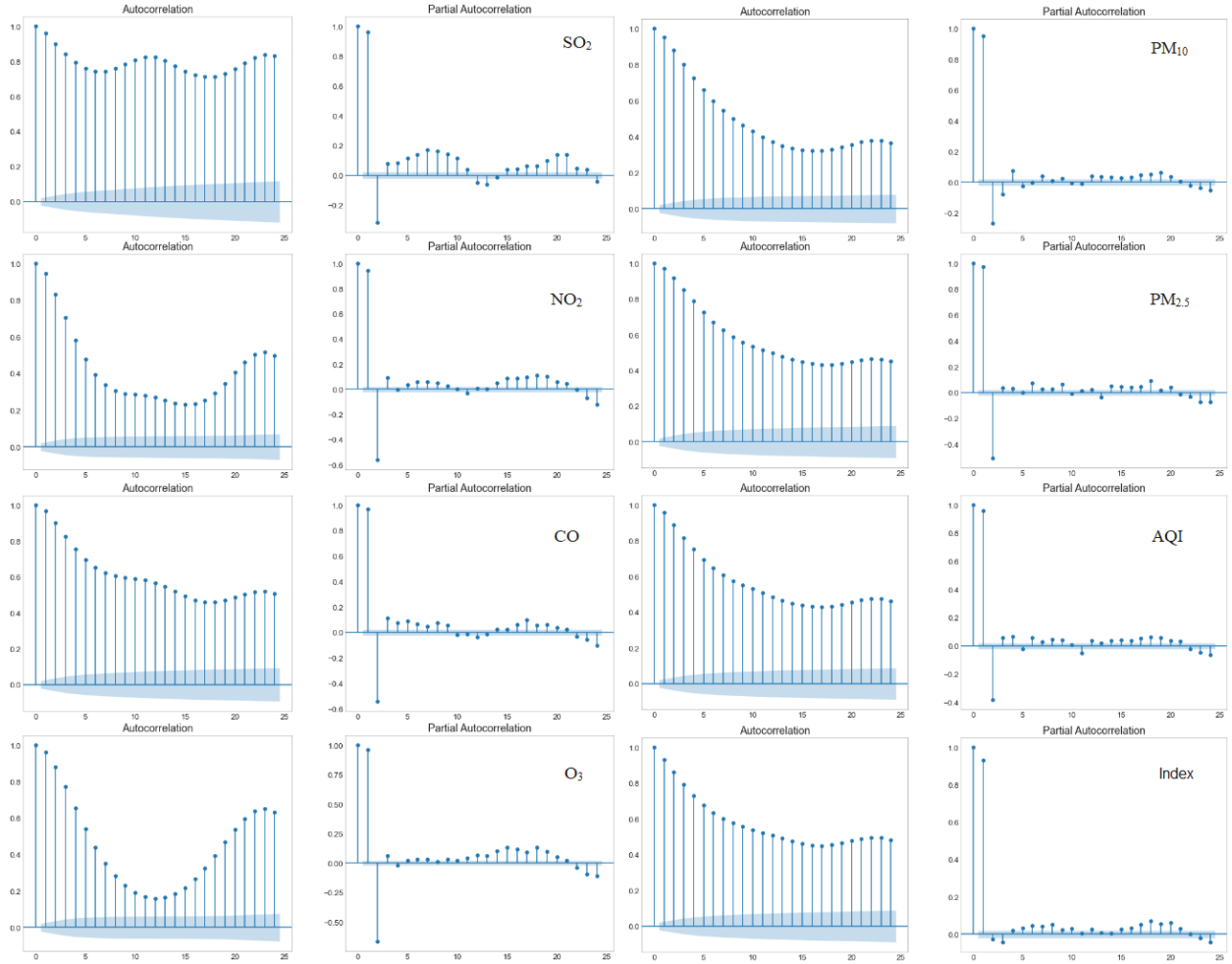


Figure 3: Preprocessing the data and determining the optimal combination of inputs

Using ACF and PACF subplots, it is possible to identify lags that are significant with the 95% confidence interval for each pollutant. A lag was considered significant if its corresponding spike exceeded the 95% confidence interval, calculated as $\pm 1.96/\sqrt{n}$, where n is the sample size. This threshold is visually indicated by the shaded region in the plots. Only the lags that crossed this boundary were selected as inputs. For instance, CO exhibited significant partial autocorrelations up to lag 8, while $PM_{2.5}$ showed significance only at earlier lags. Finally, by combining the values of the pollutants corresponding to the significant lag numbers indicated by Q1, Q2, ..., and Qn, 4 different input combinations were identified separately for each pollutant, indicated by M1, M2, M3, and M4. Table 4 displays different input combinations for each pollutant. In this table, the index related to Q displays the lag number.

Table 4: Input combinations

	SO2	NO2	CO	O3	PM10	PM2.5	AQI
M1	Q1, Q2, Q3, Q4	Q1, Q2, Q3, Q5	Q1, Q2, Q3, Q4	Q1, Q2, Q3, Q11	Q1, Q2, Q3, Q4	Q1, Q2, Q6, Q9	Q1, Q2, Q3, Q4
M2	Q1, Q2, Q3, Q4, Q5	Q1, Q2, Q3, Q5, Q6	Q1, Q2, Q3, Q4, Q5	Q1, Q2, Q3, Q11, Q12	Q1, Q2, Q3, Q4, Q7	Q1, Q2, Q6, Q9, Q13	Q1, Q2, Q3, Q4, Q6
M3	Q1, Q2, Q3, Q4, Q5, Q6	Q1, Q2, Q3, Q5, Q6, Q7	Q1, Q2, Q3, Q4, Q5, Q6	Q1, Q2, Q3, Q11, Q12, Q13	Q1, Q2, Q3, Q4, Q7, Q12	Q1, Q2, Q6, Q9, Q13, Q14	Q1, Q2, Q3, Q4, Q6, Q8
M4	Q1, Q2, Q3, Q4, Q5, Q6, Q7	Q1, Q2, Q3, Q5, Q6, Q7, Q8	Q1, Q2, Q3, Q4, Q5, Q6, Q8	Q1, Q2, Q3, Q11, Q12, Q13, Q14	Q1, Q2, Q3, Q4, Q7, Q12, Q13	Q1, Q2, Q6, Q9, Q13, Q14, Q15	Q1, Q2, Q3, Q4, Q6, Q8, Q9

In the following, the SVM method was used to select the best input combinations. Table 5 displays the values of R2 and RMSE indices in input combinations.

Table 5: Select the best input combinations

	SO2		NO2		CO		O3		PM10		PM2.5		AQI	
	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
M1	0.8641	12.238	0.8953	8.5779	0.8671	0.2564	0.90741	6.89254	0.9091	42.8492	0.7947	65.7932	0.9593	21.276
M2	0.8691	11.807	0.9063	8.1550	0.8696	0.2528	0.93155	6.05131	0.9172	37.6571	0.8075	63.8989	0.9696	19.880
M3	0.8703827676	11.730948 6	0.9027834343	8.4090570 2	0.8691893535	0.2510394242	0.9341055	5.86791177	0.9082492121	43.076434 55	0.8133495121	62.587421 22	0.9656406 26	21.318339 29
M4	0.8681116969	11.636827 03	0.9065246544	8.2183704 4	0.8702782483	0.2524283	0.9345599	5.91842516	0.9041913535	44.136486 03	0.8240649797	61.810775 66	0.9648719 4	22.056339 74

According to the findings of this table, for SO2, the values of R2, and RMSE indices related to M3 were better than the corresponding values in other input combinations. Therefore, M3 was chosen as the best input combination. For all NO2, CO, O3, and PM2.5 pollutants, M4 was selected as the best input combination. Also, for AQI, M2 was selected as the best input combination.

4 Results related to the prediction of pollutants

This section looks at and analyzes the pollution forecasting findings using hybrid frameworks. A time record of the actual and anticipated pollution levels based on the hybrid frameworks (CatBoost-AOA and CatBoost-HGS) divided by test and train databases is presented in Figure 4. In this figure, the black line represents the actual observed values (Target), while the green and yellow lines indicate the predicted values generated by the CatBoost-HGS and CatBoost-AOA models, respectively. Each subplot corresponds to a specific pollutant (AQI, CO, NO_2, O_3, PM_{10}, $PM_{2.5}$, SO_2), illustrating the temporal variations captured by the models. It is evident that all hybrid frameworks' time series curves suit the real data's time series curve rather well. The close alignment between predicted and actual values confirms the reliability of these hybrid models in accurately estimating pollutant concentrations. Additionally, it can be observed that CatBoost-HGS exhibits stronger predictive accuracy for AQI and CO, whereas CatBoost-AOA provides improved performance in pollutants with complex fluctuations, such as O_3 and PM_{10}. This demonstrates that the observational data have been properly estimated by all ML frameworks in both the test and train intervals.

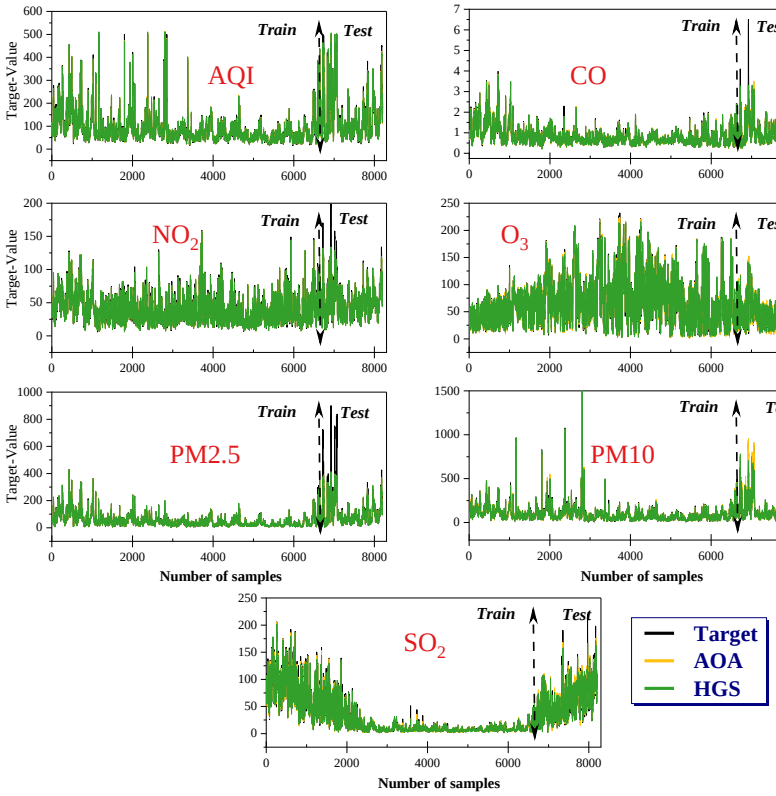
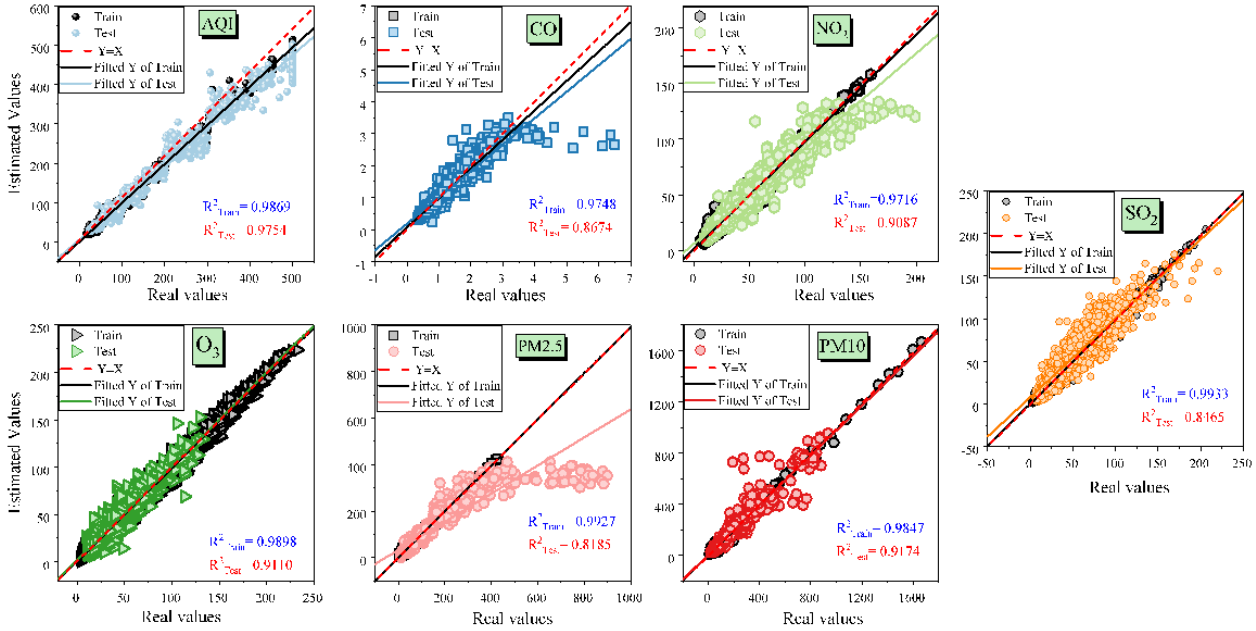
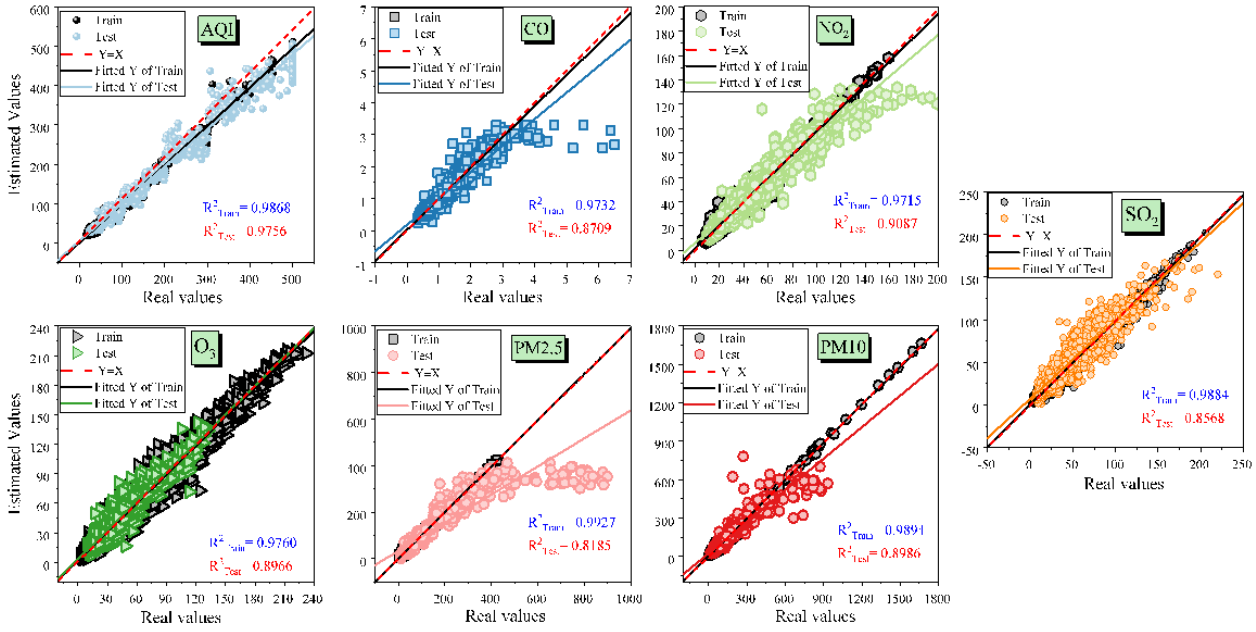


Figure 4: Time series of the actual and projected values based on the hybrid frameworks

The following will include a comparison of the frameworks based on the assessment indicators to offer a more realistic comparison of the hybrid frameworks. The hybrid framework's observation-prediction scatter plot, divided by the train and test databases, is displayed in Figure 5. The hybrid framework's observation-prediction scatter plot, divided by the train and test databases, is displayed in Figure 5. The R2 index is also displayed for each database. Based on the training database, this figure displays that both hybrid frameworks' performance is appropriate for all contaminants. The CatBoost-HGS hybrid framework demonstrated higher R2 values in the prediction of CO and AQI contaminants based on the test database. Therefore, for these 2 pollutants, the CatBoost-HGS hybrid framework is more accurate. R2 values, however, indicate that the CatBoost-AOA hybrid framework performs better in predicting PM10 and O3 contaminants drawing on the test database. Furthermore, the analysis of R2 values using the test database demonstrates that both hybrid frameworks predict NO2 and PM2.5 pollutants with excellent accuracy and near equality.



(a) CatBoost-AOA hybrid framework



(b) CatBoost-HGS hybrid framework

Figure 5: The observation-prediction scatter plot

This section examines the productivity of hybrid frameworks by concentrating on the values of the statistical assessment indicators. The values of the statistical evaluation indices for the CatBoost-AOA and

CatBoost-HGS hybrid frameworks for the test and train databases are displayed in Tables 6 and 7, respectively. The train and test databases were used to construct these indexes independently.

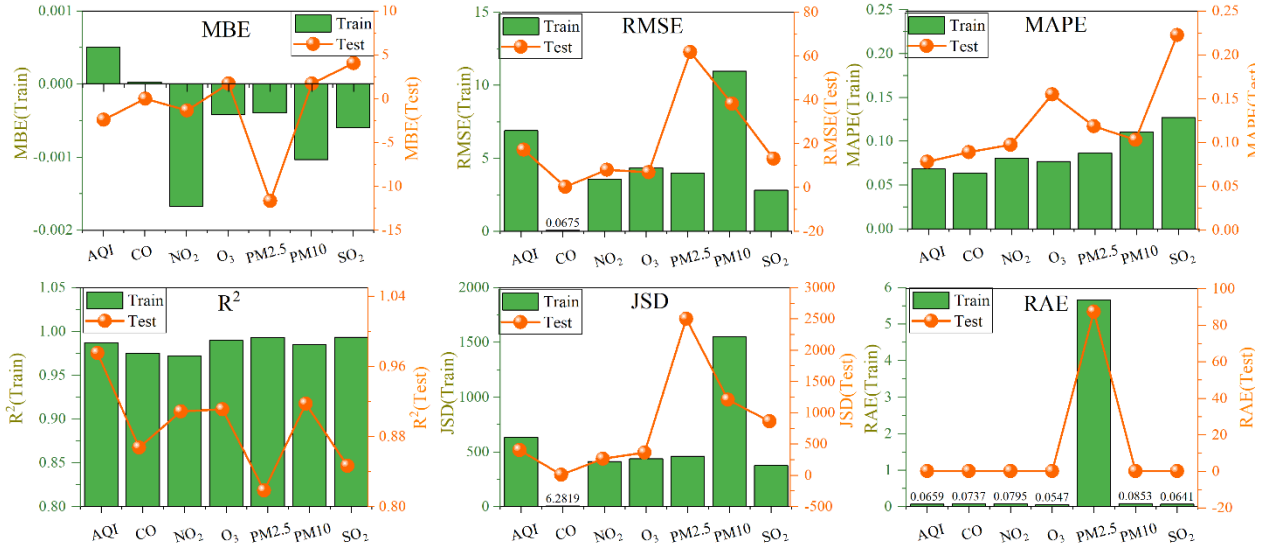
Table 6: The statistical evaluation indices related to the CatBoost-AOA hybrid framework

	AQI	CO	NO2	O3	PM2.5	PM10
	Train					
MBE	0.000504	2.49E-05	-0.00167	-0.00042	-0.0004	-0.00104
RMSE	6.896245	0.067492	3.572921	4.345468	3.999833	10.9663
MAPE	0.068246	0.063471	0.080875	0.076767	0.086226	0.110219
R2s	0.98688	0.974821	0.971562	0.98982	0.992733	0.984699
JSD	631.628	6.281885	412.0346	436.8669	460.9586	1548.496
RAE	0.065875	0.073739	0.079478	0.054661	5.657481	0.085293
	Test					
MBE	-2.40123	-0.01414	-1.36256	1.705325	-11.676	1.708281
RMSE	17.2221	0.25353	8.029421	6.933788	61.76703	38.2423
MAPE	0.077989	0.088881	0.097297	0.154836	0.118598	0.102993
R2s	0.975361	0.867402	0.908667	0.911033	0.818498	0.91736
JSD	404.9435	7.678394	265.9971	363.296	2502.604	1208.524
RAE	0.098815	0.190031	0.138356	0.156295	87.40505	0.196445

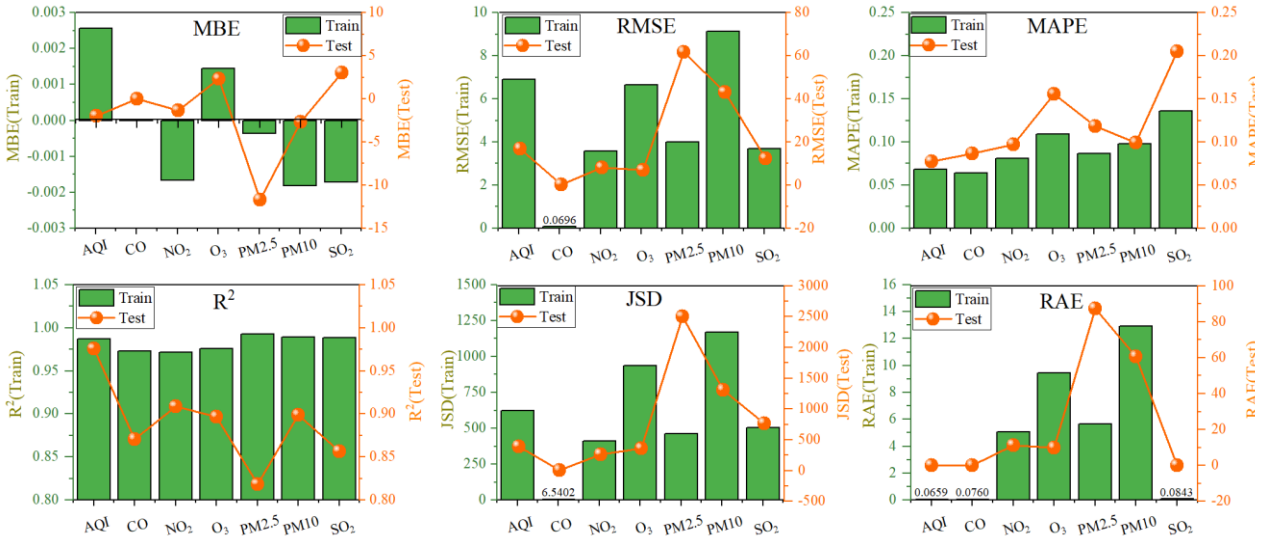
Table 7: The statistical evaluation indices related to the CatBoost-HGS hybrid framework

	AQI	CO	NO2	O3	PM2.5	PM10
	Train					
MBE	0.002563	0.000002	-0.00167	0.001449	-0.00037	-0.00181
RMSE	6.89878	0.069591	3.572889	6.663152	3.999854	9.119149
MAPE	0.068145	0.064045	0.080874	0.108812	0.086225	0.097581
R2s	0.98687	0.973229	0.971563	0.976024	0.992733	0.989421
JSD	625.0642	6.540241	412.0287	936.7454	460.963	1170.28
RAE	0.065899	0.076032	5.053598	9.424557	5.65751	12.89839
	Test					
MBE	-1.99532	-0.01647	-1.36254	2.317816	-11.676	-2.70887
RMSE	16.88679	0.250533	8.029452	6.983025	61.77031	42.97614
MAPE	0.077577	0.086798	0.097297	0.155553	0.118599	0.09913
R2s	0.975609	0.870867	0.908666	0.896567	0.818464	0.898578
JSD	393.458	7.554049	265.9998	357.9991	2502.871	1312.712
RAE	0.096891	0.187784	11.36229	9.881513	87.40969	60.81451

To make it easier to compare hybrid frameworks, Fig. 6 displays the bar plot related to the evaluation indices separately by the type of data. In the following, these indices will be examined.



(a) CatBoost-AOA hybrid framework



(b) CatBoost-HGS hybrid framework

Figure 6: Performance of hybrid frameworks based on the evaluation indices.

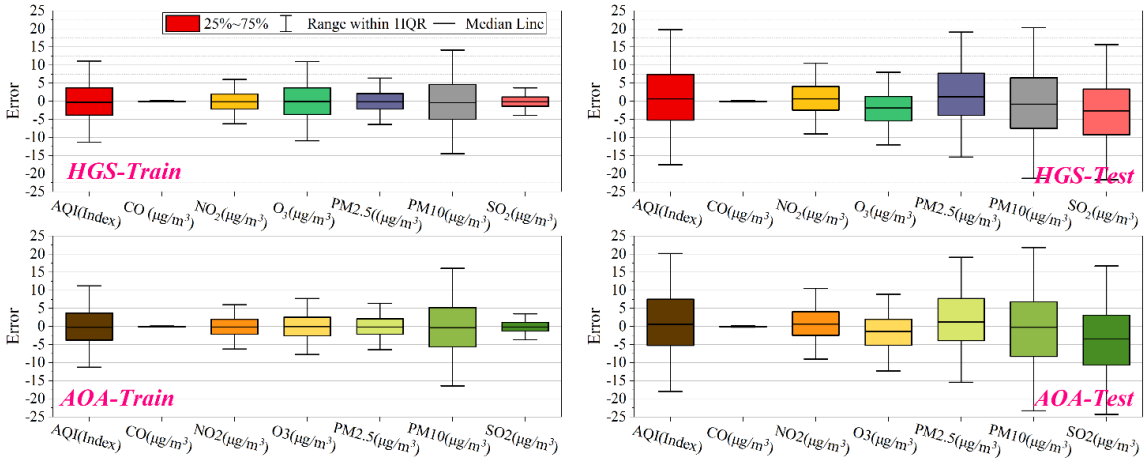


Figure 7: The box plot diagram of errors

The comparison of evaluation indices based on the outcomes of Tables 5 and 6 and Fig. 6 displays that the recommended hybrid frameworks have good accuracy in predicting pollutants. Based on the test database, it can be found that the CatBoost-HGS hybrid framework has more accurate evaluation index values in predicting AQI and CO pollutants. This is while the evaluation indicators show that the CatBoost-AOA hybrid framework is more accurate in predicting O3 and PM10 pollutants. Regarding NO2 and PM2.5 pollutants, both hybrid frameworks have the same and close evaluation index values. Therefore, both recommended hybrid frameworks have appropriate and acceptable accuracy in predicting these 2 pollutants.

The box plot diagram of errors for all hybrid frameworks is displayed in Fig. 7 based on the kind of database, i.e., train and test.

This figure illustrates how the productivity of the two suggested hybrid frameworks is nearly identical throughout both the training and testing stages. Stated differently, the locations of the box plots for each pollutant in the training and test databases are nearly identical. This figure further demonstrates the accuracy of the proposed frameworks in predicting CO concentrations. The scholars. Put otherwise, the lowest and maximum amounts of error, together with the first, second, and third quartile values, are lower than the others. Therefore, it makes sense that this model will have less inaccuracy. Moreover, the overall distribution of errors remains centered around zero, indicating the absence of systemic bias in both frameworks. The slight variations in error spread across different pollutants, particularly in O_3 and PM_{10}, suggest that while the models perform well across most cases, pollutant-specific atmospheric behaviors contribute to variations in prediction accuracy. The CatBoost-HGS model exhibits a narrower interquartile range for CO and AQI, reinforcing its stability in these predictions. In contrast, the CatBoost-AOA model shows a slightly wider error spread, likely due to its broader optimization search, which, while enhancing performance in some cases, introduces greater variance in others. These results confirm that both models provide robust predictions while maintaining a balanced trade-off between accuracy and generalization. These graphic displays that the largest plot sizes and, consequently, the largest.

The bar charts showing the overall run time for each hybrid framework are displayed in Fig. 8. This figure displays that the overall run time of the CatBoost-AOA hybrid framework is longer than the total run time of the CatBoost-HGS hybrid framework for all pollutants. As a result, the hybrid CatBoost-HGS model completes tasks more quickly. Also, using the CatBoost-AOA hybrid framework to anticipate O3 pollutants has the highest total run time among all pollutants.

This analysis highlights the trade-off between accuracy and computational cost, as CatBoost-AOA's broader optimization strategy increases runtime while potentially improving prediction performance for certain pollutants. Despite its longer execution time, CatBoost-AOA remains feasible for offline analysis, whereas CatBoost-HGS, with its lower computational burden, is more suitable for real-time applications. The feasibility of these models in real-time scenarios can be further enhanced through techniques such as parallel processing, feature selection, and model compression to optimize runtime without compromising prediction accuracy.

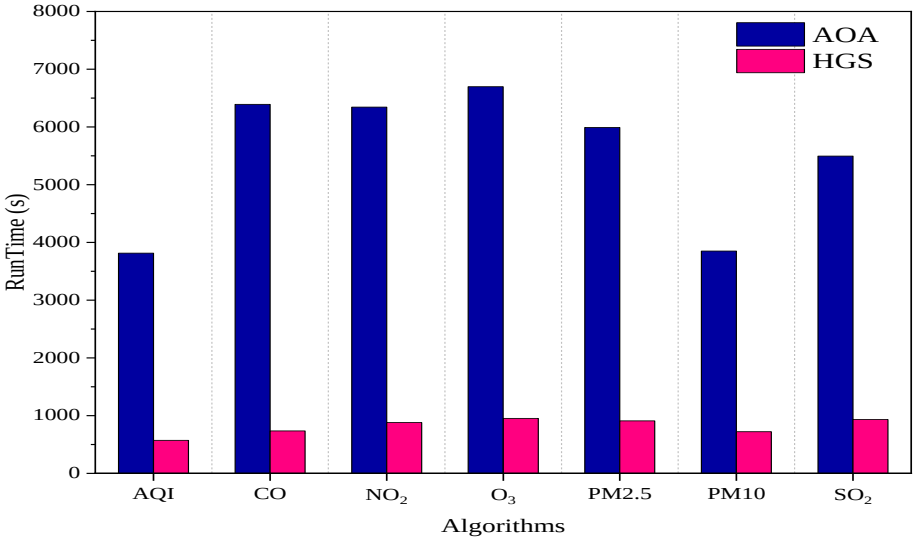


Figure 8: Comparison of the run times of hybrid frameworks

figure 9 illustrate the feature importance for various air quality prediction models, highlighting the significance of different pollutants and their influence on predicting AQI. The charts depict the relative importance of pollutants such as PM2.5, PM10, O3, CO, and AQI, with bars representing their impact on model accuracy. Figure 1 shows the ranking of various pollutants in terms of their contribution to AQI prediction. Figure 2 compares the importance of NO2 and SO2 with other pollutants, illustrating their varying effects across different features. Figure 3 presents a combined analysis of AQI, PM2.5, and PM10, showing how these pollutants together influence the overall prediction model. These visualizations help in understanding the key factors that drive air quality forecasting, assisting in more effective model optimization and feature selection.

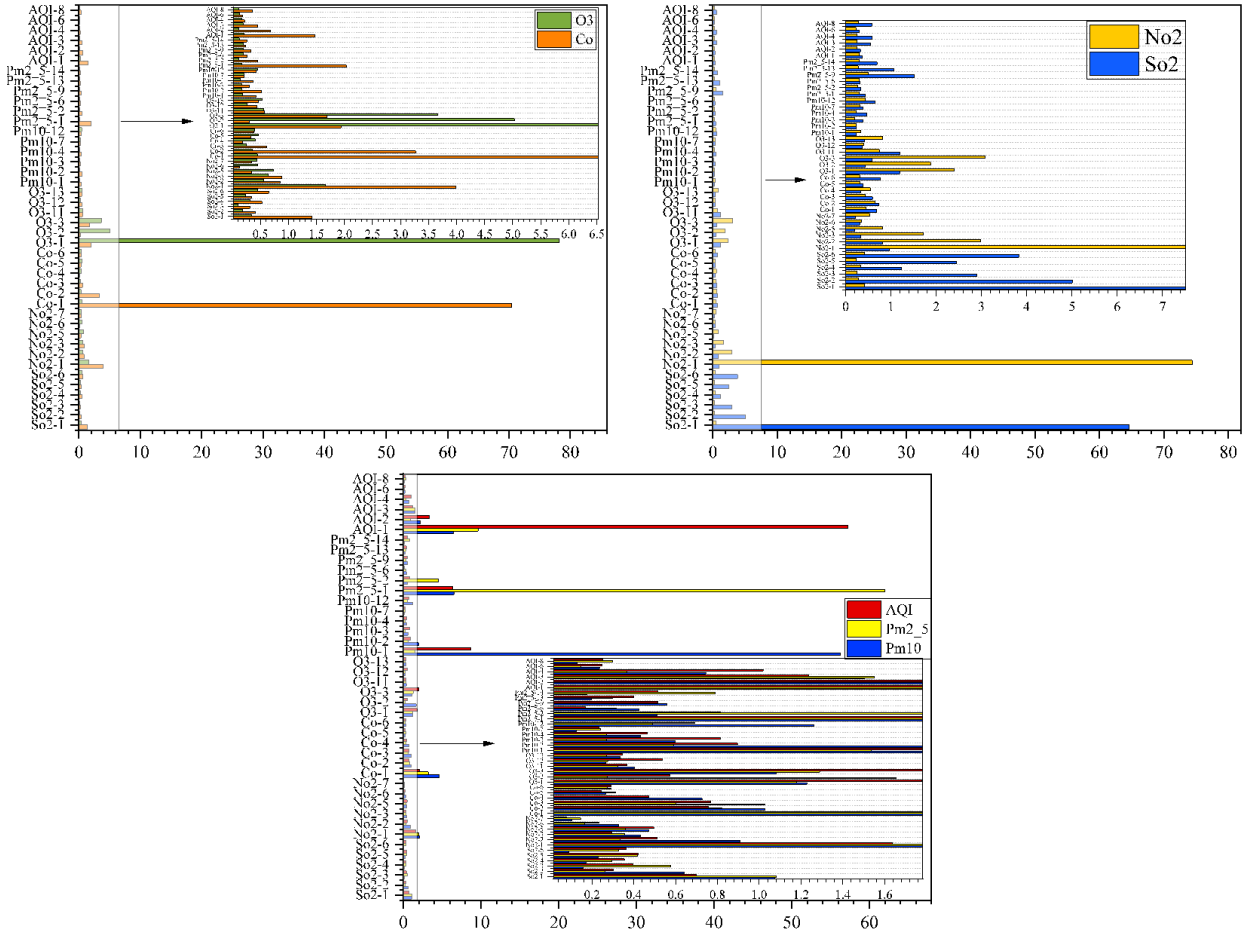


Figure 9: Feature importance analysis for air quality prediction models

Table 8 presents the results of the Wilcoxon signed-rank test to compare the performance of the AE_HGS and AE_AOA models for different pollutants. The Median Difference between the two models is provided, along with the statistical test statistic and p-value. The p-value is used to determine whether the difference is statistically significant at a significance level of $\alpha = 0.05$. If the p-value is less than 0.05, the difference is considered statistically significant. The conclusion column summarizes the outcome for each target variable, indicating whether one model performed better in terms of lower AE (Absolute Error). The Target refers to the specific pollutant or air quality index being tested (e.g., AQI, CO, NO2), and N represents the number of samples used in the test. The Compare column indicates the median difference between AE_HGS and AE_AOA, while the Statistic shows the test statistic calculated for the Wilcoxon signed-rank test. The P-value column reflects the p-value associated with the test, determining the statistical significance of the difference. Significant (α=0.05) shows whether the result is statistically significant, and the Conclusion provides a final judgment based on the p-value, indicating which model performed better in terms of AE.

Table 8: Statistical significance test results for model comparison (wilcoxon signed-rank test)

Target	N_Compared	Median_Difference (AE_HGS - AE_AOA)	Statistic	P-value	Significant (α=0.05)	Conclusion
AQI	1641	0.0072	671061.0	0.8935	FALSE	NoSignificant Difference
CO	1641	-0.0011	643982.0	0.1225	FALSE	No Significant Difference
NO2	1641	0.0000	506085.0	2.617e-18	TRUE	AOA Significantly Better (Lower AE)
O3	1641	-0.0070	672465.0	0.9516	FALSE	No Significant Difference

PM2.5	1641	-0.0000	604140.0	0.0003	TRUE	HGS Significantly Better (Lower AE)
PM10	1641	-0.3864	615175.0	0.0023	TRUE	HGS Significantly Better (Lower AE)
SO2	1641	-0.4735	592792.0	2.547e-05	TRUE	HGS Significantly Better (Lower AE)

5 Results related to AQI category classification

This section examines the CatBoost-AOA and CatBoost-HGS hybrid frameworks' accuracy in classifying the AQI. As was previously noted, the AQI is categorized into six groups according to air quality ratings. The evaluation index values for hybrid classification frameworks are displayed in Fig. 10. These graphic displays that both hybrid frameworks of classification have acceptable class classification accuracy. All of the evaluation indices for both hybrid frameworks have values of 1 based on the training database. This displays that the frameworks' training data accuracy is faultless.

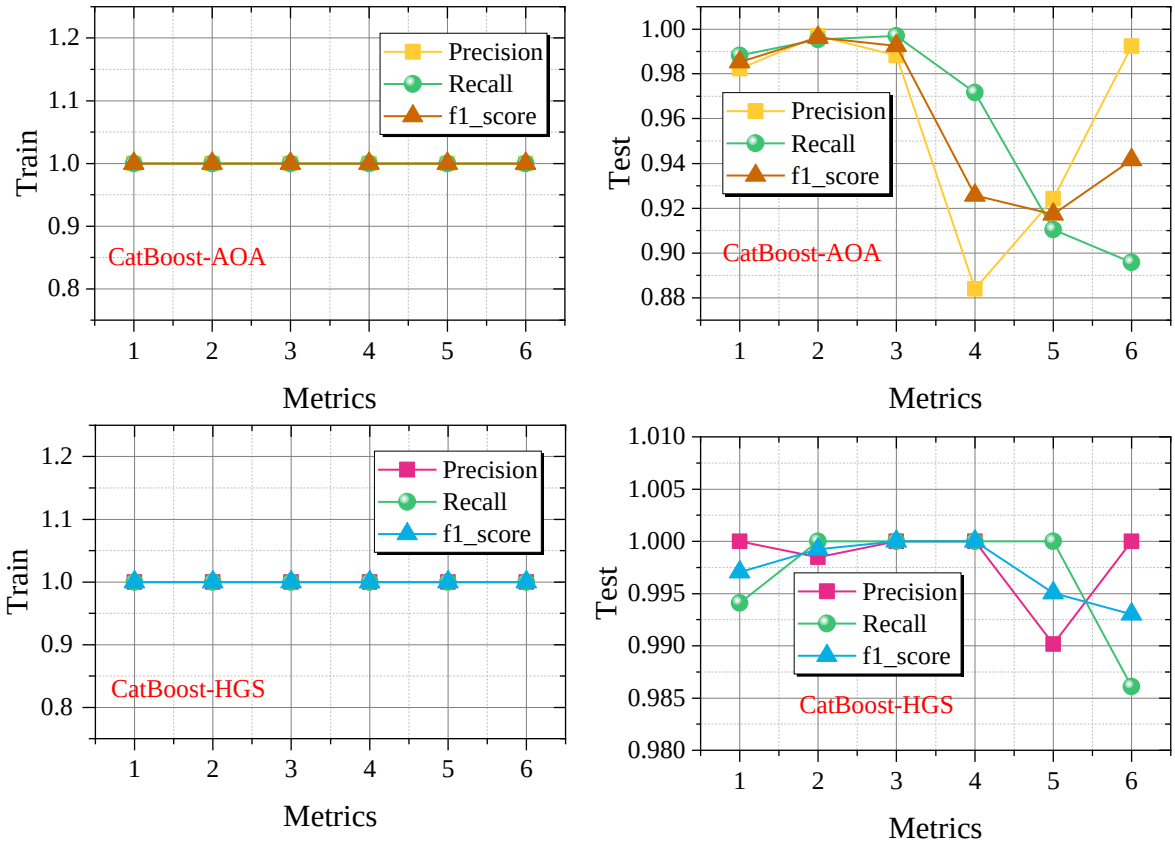
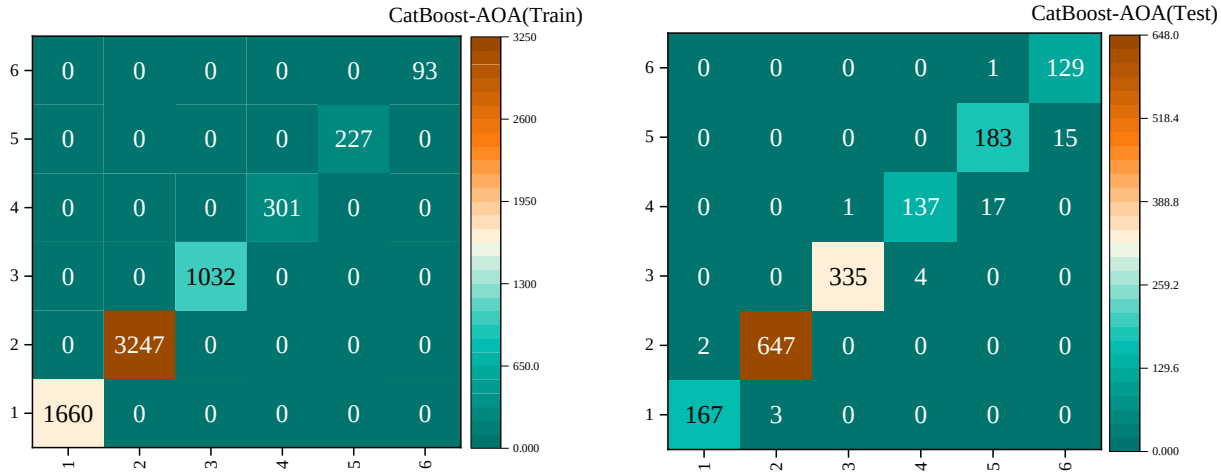
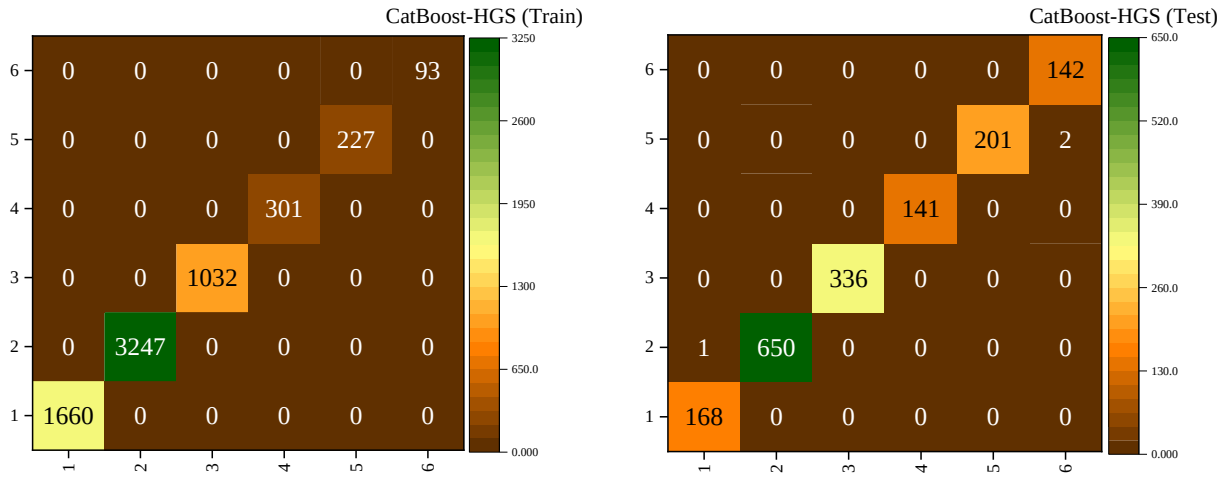


Figure 10: Comparing the accuracy of hybrid classification frameworks

The CatBoost-HGS hybrid framework's Precision, Recall, and F1 Score indices are all more than 0.985 across all classes when compared to the test database. In every class, these values are almost greater than the equivalent values of the hybrid CatBoost-AOA model. Consequently, it may be said that the hybrid CatBoost-HGS model classifies classes more accurately. The confusion matrix associated with the hybrid classification frameworks is displayed separately from the database in Fig. 11 to more precisely compare the accuracy of the classification frameworks. The horizontal axis in this picture displays the true and observed label, while the vertical axis displays the anticipated label.



(a) CatBoost-AOA hybrid framework



(b) CatBoost-HGS hybrid framework

Figure 11: The confusion matrix related to the hybrid classification frameworks

This image displays that both hybrid frameworks were able to be fully and accurately trained based on observational classifications using the training database. The values of the primary diameter cells in all classes for the CatBoost-HGS hybrid framework are greater than the comparable values in the CatBoost-AOA hybrid framework, according to the test database. This demonstrates the CatBoost-HGS hybrid framework's increased accuracy in classifying data using the test database.

6 Discussion

The results demonstrate that the proposed CatBoost-HGS model outperforms traditional models like XGBoost (Song et al., 2023) and SARIMA (Bhatti et al., 2021) in most pollutants, particularly CO and AQI. As seen in Table 6, CatBoost-HGS achieves an R^2 of 0.9756 for AQI and 0.870867 for CO, reflecting strong predictive capabilities. Compared to SARIMA, which struggles with nonlinear dependencies, and XGBoost, which lacks dynamic hyperparameter tuning, CatBoost-HGS leverages the Hunger Games Search (HGS) algorithm for superior parameter optimization, leading to more precise predictions. However, a comparative analysis with CatBoost-AOA in Table 5 reveals that while CatBoost-HGS excels in CO and AQI prediction, it slightly underperforms in O_3 and PM_{10} forecasting. Specifically, CatBoost-HGS achieves an RMSE of 0.250533 for CO and 16.88679 for AQI, whereas CatBoost-AOA records slightly higher errors (0.25353 and 17.2221, respectively). However, for O_3 and PM_{10}, CatBoost-AOA performs better with an R^2 of 0.911033 for O_3 compared to 0.896567 in HGS. This can be attributed to AOA's stronger global search capability, which is more suited for pollutants with complex temporal and chemical interactions. Conversely, HGS excels in local optimization, making it more effective for pollutants with clearer short-term patterns like CO.

The analysis also confirms that classification-based models yield superior AQI predictions compared to regression-based approaches. Given AQI's discrete categorization (e.g., Good, Moderate, Unhealthy), classification models like CatBoost-HGS (R^2 = 0.9756 for AQI) better capture these structured thresholds than regression models, which may misinterpret transitions between AQI classes. This highlights the advantage of

classification in enhancing interpretability and aligning with air quality management needs.

However, this observation is dataset-dependent, and while classification demonstrated superior performance in this study, its generalizability to different datasets should be further explored. Incorporating additional air quality datasets from different regions could help validate the robustness of classification models over regression in AQI forecasting.

Furthermore, the sensitivity analysis (Figure 2) confirms that PM_{10} has the highest impact on AQI prediction, aligning with environmental studies highlighting its prolonged atmospheric presence and regulatory significance in AQI computation. PM_{10} is known to have significant adverse health effects and is a primary component in air quality standards worldwide. The dominance of PM_{10} in the feature importance rankings supports the model's reliability in identifying key air quality indicators. Future studies should examine whether this trend holds across diverse geographic regions and pollution sources.

One concern with the proposed models is the higher computational cost of CatBoost-AOA, which exhibits a longer runtime due to its more exhaustive optimization process. While CatBoost-AOA slightly improves accuracy in O_3 and PM_{10}, the added computational burden must be justified. To balance efficiency and accuracy, future work should explore feature selection to reduce dimensionality, parallelized processing (GPU acceleration) to speed up training, and dynamic hybrid approaches where AOA is selectively applied to pollutants that require extensive optimization.

Overall, the findings highlight that CatBoost-HGS is highly effective for AQI and CO prediction, offering a balance between accuracy and efficiency, while CatBoost-AOA is preferable for pollutants with complex atmospheric behaviors. Future research should focus on hybrid frameworks that dynamically switch between optimization techniques to further enhance predictive performance and improve computational efficiency for large-scale applications.

7 Conclusion

Environmental pollution is one of the major problems of developing countries, which has been increased by the expansion of urbanization and excessive consumption of fossil fuels. One of the biggest and most dangerous issues facing modern society is air pollution, which is made worse by the growing number of automobiles on the road. Every day, the quality of the air varies. The series of data related to air quality has random, irregular, and unstable characteristics, and this has made it difficult to anticipate pollutants and air quality. Considering the adverse consequences of air pollution on the health of persons and the environment, it is vital to lessen and address this problem based on the precise knowledge of pollutants and the factors affecting them and identifying the polluted areas; therefore, using mathematical frameworks in the form of ML is an optimal and cost-effective approach for modeling air pollution. Therefore, in this investigation, using the capabilities of hybrid frameworks based on ML, hybrid frameworks were presented to anticipate air quality pollutants and classify air quality. The recommended hybrid frameworks were created by combining the CatBoost algorithm with the AOA and HGS algorithms. The study's database includes daily time series data on China's air pollution from 2018 to 2021. Choosing the right inputs for intelligent frameworks is very important because it reduces costs, saves time, and increases the accuracy and efficiency of frameworks. To select input combinations for each pollutant, ACF and PACF approaches were used. Additionally, the SVR algorithm was employed to choose the optimal input combination for every pollutant. The accuracy of hybrid frameworks in the following data prediction and classification was compared using different evaluation indices. The results showed that both CatBoost-HGS and CatBoost-AOA hybrid frameworks have good accuracy in predicting pollutants. The outcomes demonstrated that the CatBoost-HGS hybrid framework for predicting AQI and CO pollutants and the CatBoost-AOA hybrid framework for predicting O3 and PM10 pollutants are more accurate. Regarding NO2 and PM2.5 pollutants, both hybrid frameworks have the same and close evaluation index values. Therefore, both recommended hybrid frameworks have appropriate and acceptable accuracy in predicting these 2 pollutants. Additionally, the evaluation indicators demonstrated that the CatBoost-HGS hybrid framework outperforms the CatBoost-AOA in accurately classifying and identifying the air quality class.

Nomenclature

ACF	Autocorrelation Function	PACF	Partial Autocorrelation Function
AI	Artificial Intelligence	PCA	principal component analysis
ANN	Artificial neural network	RBF	radial basis function
AOA	Arithmetic Optimization Algorithm	RF	random forest
AQI	Air quality index	SARIMA	Seasonal Autoregressive Integrated Moving Average
CatBoost	Categorical boosting	SVM	Support vector machine
DL	deep learning	SVR	Support Vector Regression
FN	False Negative	TN	True Negative
FP	False Positive	TP	True Positive

HGS	Hunger Games Search	WHO	World Health Organization
HYSPLIT	Hybrid Single-Particle Lagrangian Integrated Trajectory	XGBoost	Extreme Gradient Boosting
$JSD(P \mid\mid Q)$	Jensen-Shannon divergence between P and Q	$\hat{y}_i$	ith estimated value
LSTM	Long Short-Term Memory	$\bar{y}$	mean of data points
ML	ML	y_i	ith real value
N	number of observations		

Competing interests

The authors declare no competing interests.

Authorship contribution statement

Xiaowen GENG: Writing-Original draft preparation Conceptualization, Supervision, Project administration.

Data availability

The scholars will make the raw data supporting this article's conclusions available without undue reservation.

Conflicts of interest

The scholars claimed no conflicts of interest considering this investigation.

Author statement

The manuscript has been read and approved by all the authors, the requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript displays honest work.

Ethical approval

All scholars have been personally and actively involved in substantial work leading to the paper and will take public responsibility for its content.

References

[1] Almetwally, A.A., M. Bin-Jumah and A.A. Allam (2020). Ambient air pollution and its influence on human health and welfare: an overview. *Environmental Science and Pollution Research*, Springer Nature, 27, pp. 24815–24830. https://doi.org/10.1007/s11356-020-09042-2

[2] Ukaogo, P.O., U. Ewuzie and C. V Onwuka (2020). Environmental pollution: causes, effects, and the remedies, In *Microorganisms for Sustainable Environment and Health*, Elsevier, pp: 419–429. https://doi.org/10.1016/B978-0-12-819001-2.00021-8

[3] Zeng, Y., Y. Cao, X. Qiao, B.C. Seyler and Y. Tang (2019). Air pollution reduction in China: Recent success but great challenge for the future. *Science of the Total Environment*, Elsevier, 663, pp. 329–337. https://doi.org/10.1016/j.scitotenv.2019.01.262

[4] Manisalidis, I., E. Stavropoulou, A. Stavropoulos and E. Bezirtzoglou (2020). Environmental and health impacts of air pollution: a review. *Frontiers in Public Health*, Frontiers Media S.A, 8, pp. 14. https://doi.org/10.3389/fpubh.2020.00014

[5] Organization, W.H (2016). *Ambient air pollution: A global assessment of exposure and burden of disease*.

[6] Yang, J., R. Yan, M. Nong, J. Liao, F. Li and W. Sun (2021). PM2. 5 concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time. *Atmospheric Pollution Research*, Elsevier, 12(9), pp. 101168. https://doi.org/10.1016/j.apr.2021.101168

[7] Singh, D., M. Dahiya, R. Kumar and C. Nanda (2021). Sensors and systems for air quality assessment monitoring and management: A review. *Journal of Environmental Management*, Elsevier, 289, pp. 112510. https://doi.org/10.1016/j.jenvman.2021.112510

[8] Liao, K., X. Huang, H. Dang, Y. Ren, S. Zuo and C. Duan (2021). Statistical approaches for forecasting primary air pollutants: a review. *Atmosphere*, MDPI, 12(6), pp. 686. https://doi.org/10.3390/atmos12060686

[9] Kattenborn, T., J. Leitloff, F. Schiefer and S. Hinz (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, 173, pp. 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

[10] Song, Y., C. Zhang, X. Jin, X. Zhao, W. Huang, X. Sun, Z. Yang and S. Wang (2023). Spatial prediction of PM2. 5 concentration using hyper-parameter optimization XGBoost model in China. *Environmental Technology & Innovation*, Elsevier, 32, pp. 103272. https://doi.org/10.1016/j.eti.2023.103272

[11] Bhatti, U.A., Y. Yan, M. Zhou, S. Ali, A. Hussain, H. Qingsong, Z. Yu and L. Yuan (2021). Time series analysis and forecasting of air pollution particulate matter (PM 2.5): an SARIMA and factor analysis approach. *IEEE Access*, IEEE, 9, pp. 41019–41031. https://doi.org/10.1109/ACCESS.2021.3060744

[12] Liu, H. and C. Chen (2020). Prediction of outdoor PM2. 5 concentrations based on a three-stage hybrid neural network model. *Atmospheric Pollution Research*, Elsevier, 11(3), pp. 469–481. https://doi.org/10.1016/j.apr.2019.11.019

[13] Kaushal, A., A.K. Gupta and V.K. Sehgal (2025). Hybrid CatBoost and SVR Model for Earthquake Prediction Using the LANL Earthquake Dataset. *Informatica*, Slovenian Society Informatika, 49(14). https://doi.org/10.31449/inf.v49i14.6524

[14] Amnuaylojaroen, T (2022). Prediction of PM2. 5 in an urban area of northern Thailand using multivariate linear regression model. *Advances in Meteorology*, WILEY Online Library, 2022(1), pp. 3190484. https://doi.org/10.1155/2022/3190484

[15] Kong, Y., L. Sheng, Y. Li, W. Zhang, Y. Zhou, W. Wang and Y. Zhao (2021). Improving PM2. 5 forecast during haze episodes over China based on a coupled 4D-LETKF and WRF-Chem system. *Atmospheric Research*, Elsevier, 249, pp. 105366. https://doi.org/10.1016/j.atmosres.2020.105366

[16] Shams, S.R., S. Kalantary, A. Jahani, S.M.P. Shams, B. Kalantari, D. Singh, M. Moeinnadini and Y. Choi (2023). Assessing the effectiveness of artificial neural networks (ANN) and multiple linear regressions (MLR) in forcasting AQI and PM10 and evaluating health impacts through AirQ+ (case study: Tehran). *Environmental Pollution*, Elsevier, 338, pp. 122623. https://doi.org/10.1016/j.envpol.2023.122623

[17] Rastgoo, A. and H. Khajavi (2023). A novel study on forecasting the airfoil self-noise, using a hybrid model based on the combination of CatBoost and Arithmetic Optimization Algorithm. *Expert Systems with Applications*, Elsevier, 229, pp. 120576. https://doi.org/10.1016/j.eswa.2023.120576

[18] Khajavi, H. and A. Rastgoo (2023). Improving the prediction of heating energy consumed at residential buildings using a combination of support vector regression and meta-heuristic algorithms. *Energy*, Elsevier, 272, pp. 127069. https://doi.org/10.1016/j.energy.2023.127069

[19] Bentéjac, C., A. Csörgő and G. Martínez-Muñoz (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, Springer Nature, 54, pp. 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[20] Luo, M., Y. Wang, Y. Xie, L. Zhou, J. Qiao, S. Qiu and Y. Sun (2021). Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. *Forests*, MDPI, 12(2), pp. 216. https://doi.org/10.3390/f12020216

[21] Drucker, H., C.J. Burges, L. Kaufman, A. Smola and V. Vapnik (1996). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9.

[22] Yu, H. and S. Kim (2012). SVM Tutorial-Classification, Regression and Ranking. *Handbook of Natural Computing*, Springer Nature, 1, pp. 479–506. https://doi.org/10.1007/978-3-540-92910-9_15

[23] Ghasemi, E., H. Kalhori and R. Bagherpour (2016). A new hybrid ANFIS–PSO model for prediction of peak particle velocity due to bench blasting. *Engineering with Computers*, Springer Nature, 32, pp. 607–614. https://doi.org/10.1007/s00366-016-0438-1

[24] Yang, Y., H. Chen, A.A. Heidari and A.H. Gandomi (2021). Hunger games search: Visions, conception, implementation, deep analysis, perspectives, and towards performance shifts. *Expert Systems with Applications*, Elsevier, 177, pp. 114864. https://doi.org/10.1016/j.eswa.2021.114864

[25] Nguyen, H. and X.-N. Bui (2021). A novel hunger games search optimization-based artificial neural network for predicting ground vibration intensity induced by mine blasting. *Natural Resources Research*, Springer Nature, 30(5), pp. 3865–3880. https://doi.org/10.1007/s11053-021-09903-8

[26] AbuShanab, W.S., M. Abd Elaziz, E.I. Ghandourah, E.B. Moustafa and A.H. Elsheikh (2021). A new fine-tuned random vector functional link model using Hunger games search optimizer for modeling friction stir welding process of polymeric materials. *Journal of Materials Research and Technology*, Elsevier, 14, pp. 1482–1493. https://doi.org/10.1016/j.jmrt.2021.07.031

[27] Abualigah, L., A. Diabat, S. Mirjalili, M. Abd Elaziz and A.H. Gandomi (2021). The arithmetic optimization algorithm. *Computer Methods in Applied Mechanics and Engineering*, Elsevier, 376, pp. 113609. https://doi.org/10.1016/j.cma.2020.113609

[28] Zhang, Y.-J., Y.-X. Yan, J. Zhao and Z.-M. Gao (2022). AOAAO: The hybrid algorithm of arithmetic optimization algorithm with aquila optimizer. *IEEE Access*, IEEE, 10, pp. 10907–10933. https://doi.org/10.1109/ACCESS.2022.3144431

[29] Kaveh, A., K.B. Hamedani and M. Kamalinejad (2021). Improved arithmetic optimization algorithm for structural optimization with frequency constraints. *Int J Optim Civil Eng*, 11(4), pp. 663–693. http://ijoce.iust.ac.ir/article-1-500-fa.html

[30] H. Zhang (2022). *Data sets of AQI, CO and NO2 and O3, PM10 and PM2.5, SO2 from 264 city-level monitoring stations in China*. https://dx.doi.org/10.21227/vjqn-3e65

[31] Khavari, B., A. Sahlberg, W. Usher, A. Korkovelos and F.F. Nerini (2021). The effects of population aggregation in geospatial electrification planning. *Energy Strategy Reviews*, Elsevier, 38, pp. 100752. https://doi.org/10.1016/j.esr.2021.100752